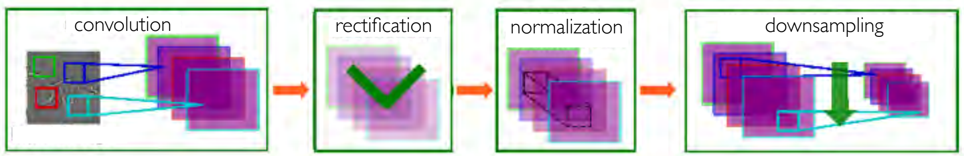
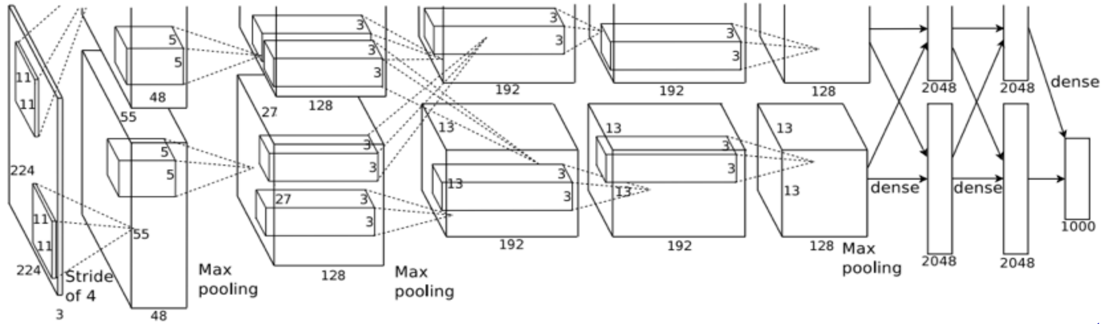
On The Landscape of Empirical Risks

1

Yuan YAO
HKUST
Based on Tomaso Poggio, Joan Bruna et al.

Empirical Risk



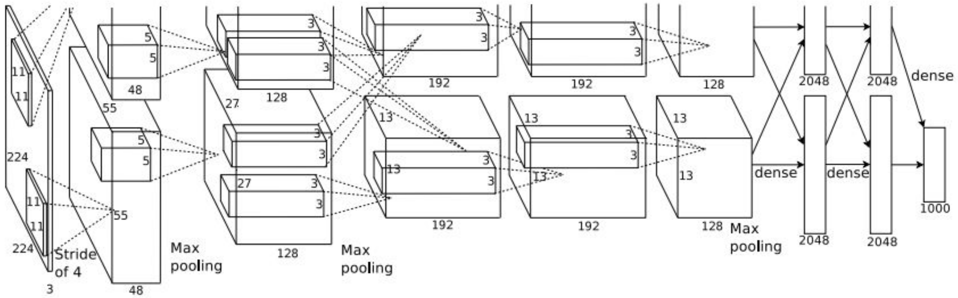


input $= x$ parameters $= w$ output $= f(x; w)$

$$R_S[w] = \frac{1}{N} \sum_{k=1}^{N} \text{loss}\left(f(x_k; w), y_k\right)$$

Case Study: AlexNet

[Krizhevsky et al. 2012]



Full (simplified) AlexNet architecture:
[227x227x3] INPUT
[55x55x96] CONV1: 96 11x11 filters at stride 4, pad 0
[27x27x96] MAX POOL1: 3x3 filters at stride 2
[27x27x96] NORM1: Normalization layer
[27x27x256] CONV2: 256 5x5 filters at stride 1, pad 2
[13x13x256] MAX POOL2: 3x3 filters at stride 2
[13x13x256] NORM2: Normalization layer
[13x13x384] CONV3: 384 3x3 filters at stride 1, pad 1
[13x13x384] CONV4: 384 3x3 filters at stride 1, pad 1
[13x13x256] CONV5: 256 3x3 filters at stride 1, pad 1
[6x6x256] MAX POOL3: 3x3 filters at stride 2
[4096] FC6: 4096 neurons
[4096] FC7: 4096 neurons
[1000] FC8: 1000 neurons (class scores)

#params =
11*11*3*96+5*5*256+384*3*3
+384*3*3+256*3*3
=50,464

#params =
6*6*256*4096+4096*4096+4096*1000=58.6M

Generalization: Population vs. Empirical Risks

Given: i.i.d. sample $S = \{z_1, \ldots, z_n\}$ from dist D

Goal: Find a good predictor function f

$$R[f] = \mathbb{E}_z \operatorname{loss}(f; z)$$

Population risk

(test error)

(Test/Validation Loss; if the loss is 0/1 indicator function, then called 'test error')

unknown!

$$R_S[f] = \frac{1}{n} \sum_{i=1}^{n} \operatorname{loss}(f; z_i)$$

Empirical risk

(training error)

(Training Loss)

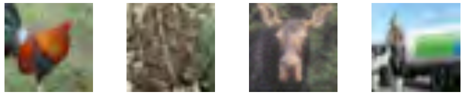
Minimize using SGD!

$$\boxed{\text{Generalization error:} \quad R[f] - R_S[f]}$$

How much empirical risk underestimates population risk

We can compute R_S...

When is it a good proxy for R?



CIFAR10

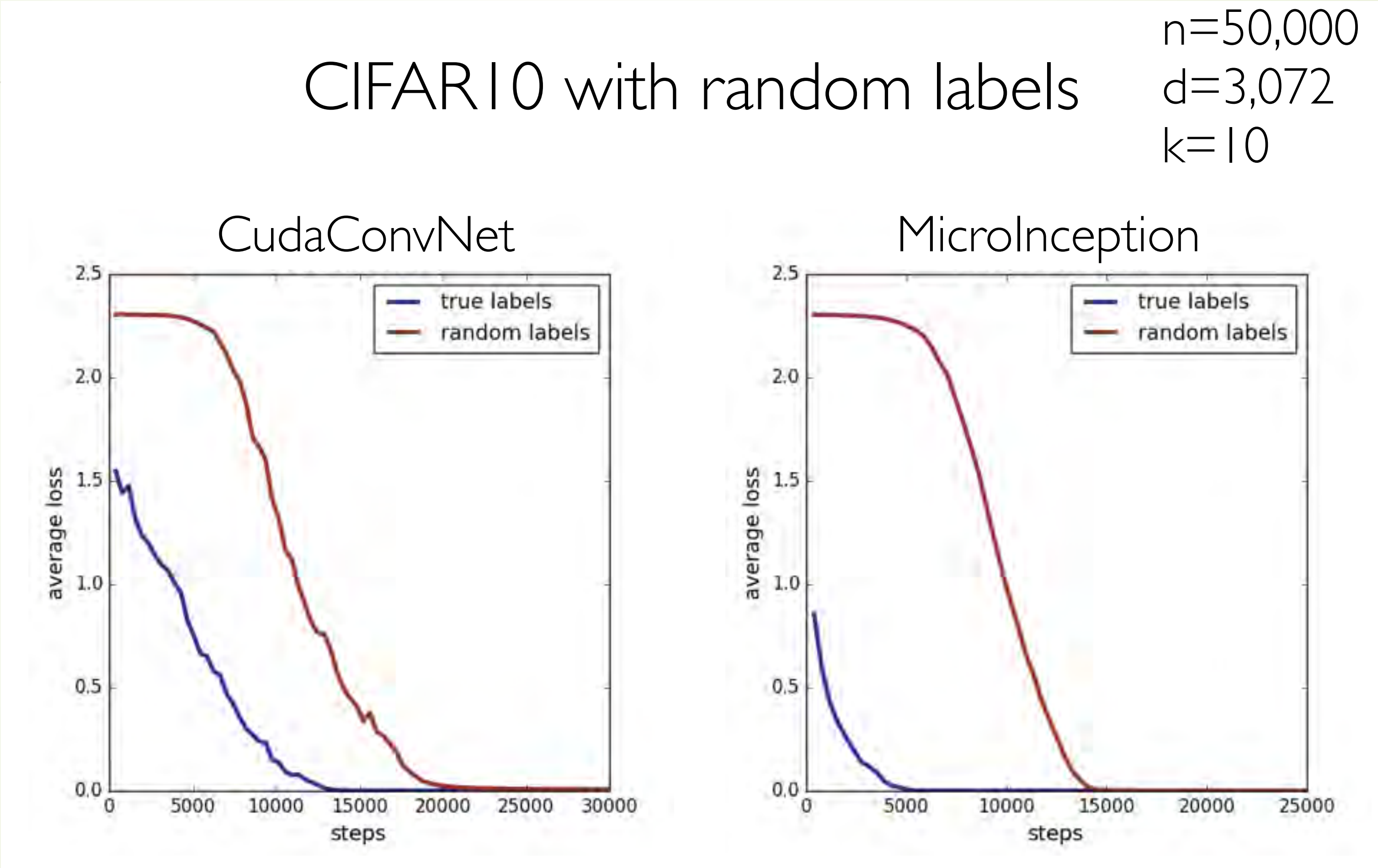
n=50,000
d=3,072
k=10

What happens when I turn off the regularizers?

Model	parameters	p/n	Train ***loss***	Test ***error***
CudaConvNet	145,578	2.9	0	23%
CudaConvNet (with regularization)	145,578	2.9	0.34	18%
MicroInception	1,649,402	33	0	14%
ResNet	2,401,440	48	0	13%

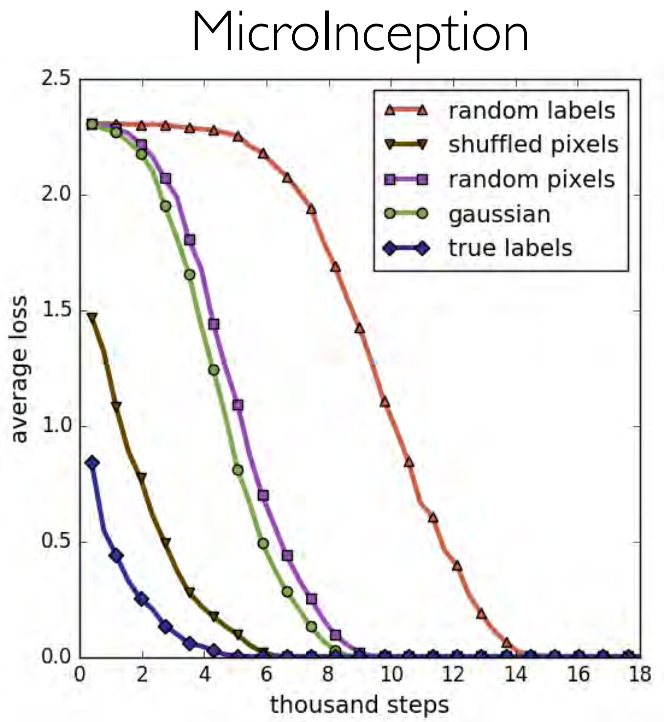
Recht 2017 FoCM.

Global optima found as zero training error



From Ben Recht 2017 FoCM

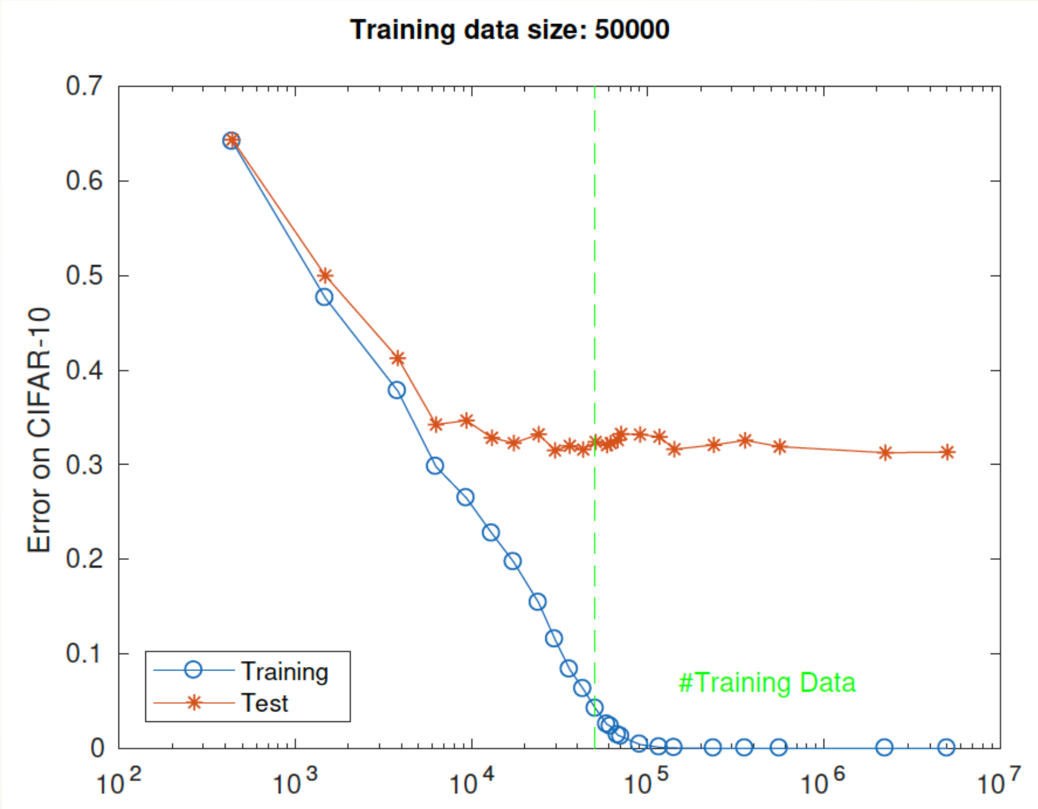
Cifar10 with randomized experiments



n=50,000
d=3,072
k=10
p=1,649,402

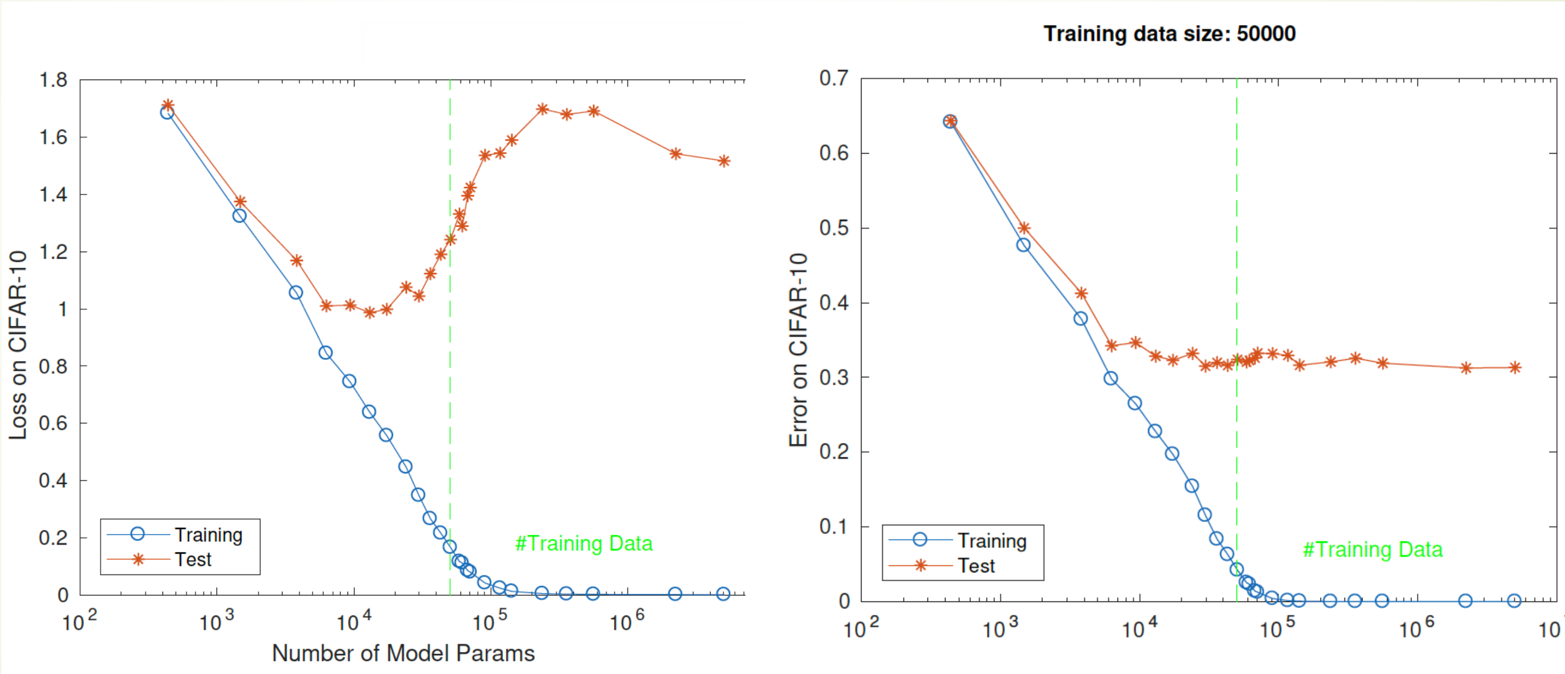
From Ben Recht 2017 FoCM: training faster, generalize better

Big models does not overfit…



Tommy Poggio, 2018

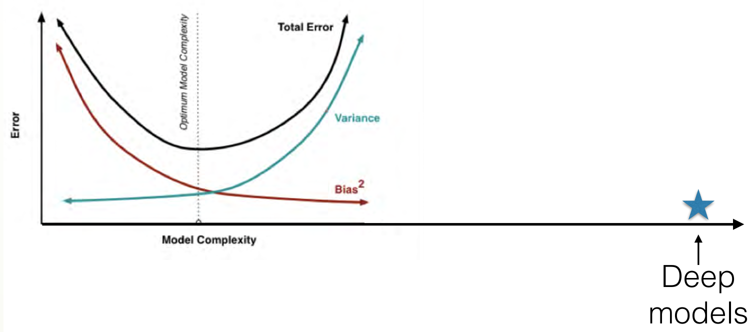
Big models may overfit test loss, but generalize well in test error



Tommy Poggio, 2018

New challenges to understanding

- Big (**overparametric**) models with SGD may find **global optima** efficiently
- Big (**overparametric**) models may **generalize** well
- Why? Possible answers:
  - Global optima of overparametric empirical risks are degenerate, favor for SGD
  - The landscape of empirical risks of overparametric models might be simple
  - Gradient based algorithms tend to find max margin models which generalize well



Models where p>20n are common

Recall: SGD behaves like Gradient Descent Langevin dynamics (SDE)

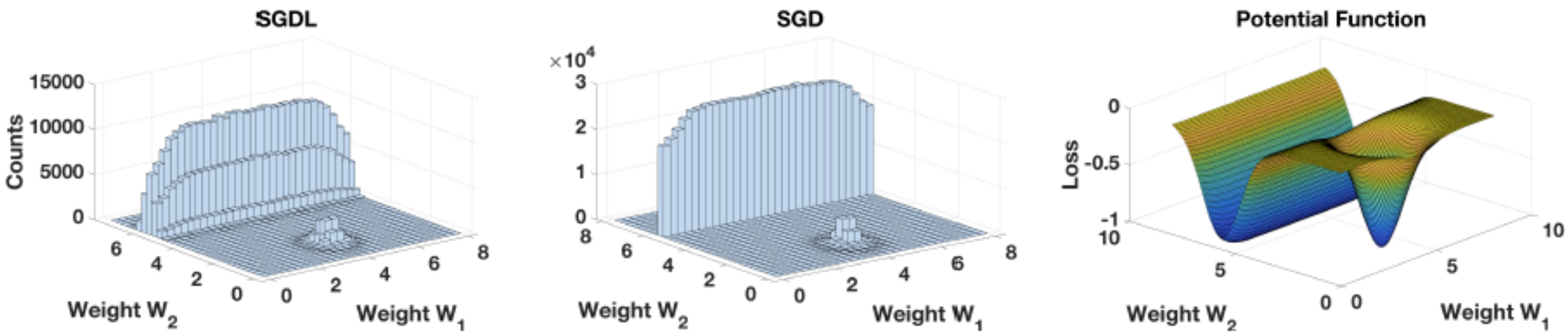
$$\frac{dw}{dt} = -\gamma_t \nabla V(w(t), z(t)) + \gamma_t' dB(t)$$

with the Boltzmann equation as asymptotic "solution"

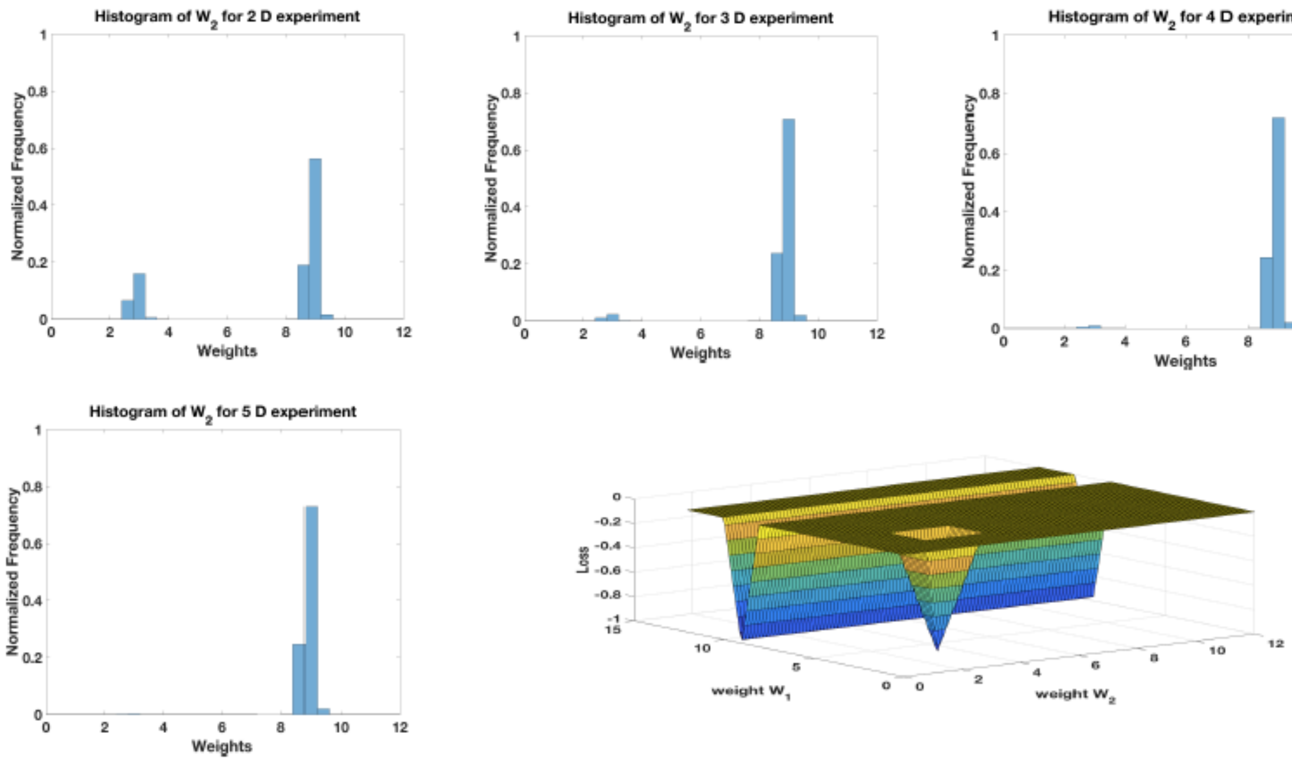
$$p(w) \sim \frac{1}{Z} = e^{-\frac{V(w)}{T}}$$

SGD/GDL selects larger volume minima e.g. degenerate

GDL ~ SGD (empirically)



Concentration because of high dimensionality



Poggio, Rakhlin, Golovitc, Zhang, Liao, 2017

Summary

- For overparametric deep networks, there are many degenerate (flat) optimizers, including the global minima

- Gradient Descent Langevin dynamics finds with overwhelming probability the flat, large volume global minima (zero-training loss), and SGD behaves in a similar way empirically

Topology and Geometry of Empirical Risk Landscapes for Multilinear and 2-Layer Rectified Networks

Based on Joan Bruna et al.

- We consider the standard ML setup:

$$\hat{E}(\Theta) = \mathbb{E}_{(X,Y)\sim\hat{P}}\ell(\Phi(X;\Theta),Y) + \mathcal{R}(\Theta)$$
$$E(\Theta) = \mathbb{E}_{(X,Y)\sim P}\ \ell(\Phi(X;\Theta),Y)\ .$$

$$\hat{P} = \frac{1}{n}\sum_{i\leq}\delta_{(xi,y_i)}$$

$\ell(z)$ convex

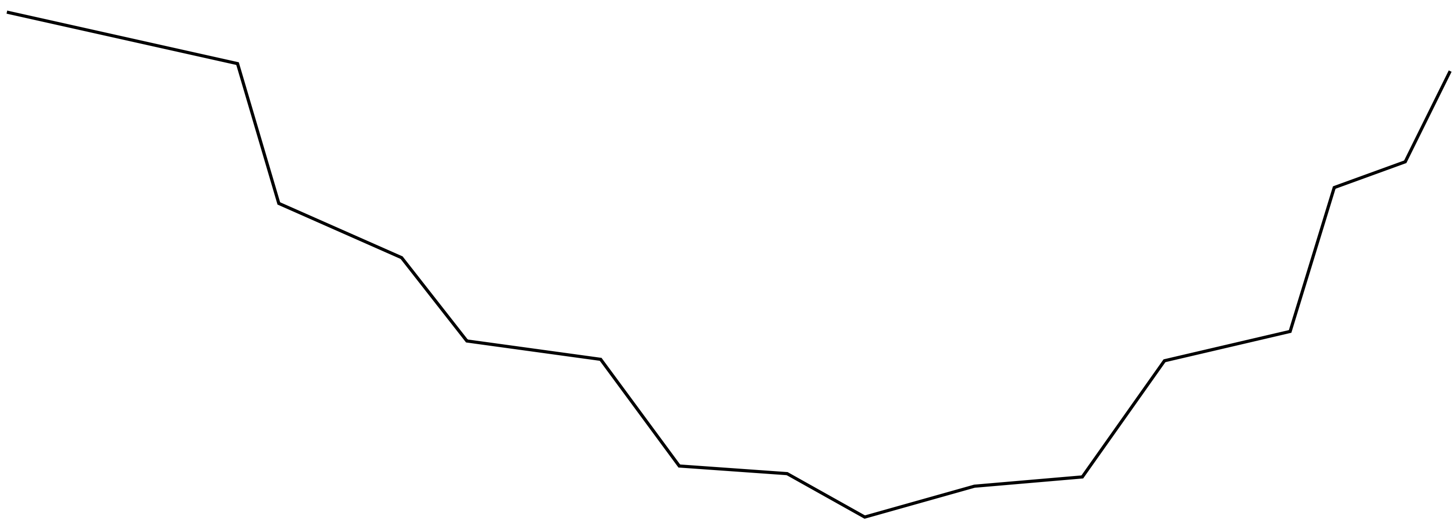
$\mathcal{R}(\Theta)$: regularization

- Population loss decomposition (*aka* "fundamental theorem of ML"):

$$E(\Theta^*) = \underbrace{\hat{E}(\Theta^*)}_{\text{training error}} + \underbrace{E(\Theta^*) - \hat{E}(\Theta^*)}_{\text{generalization gap}}\ .$$

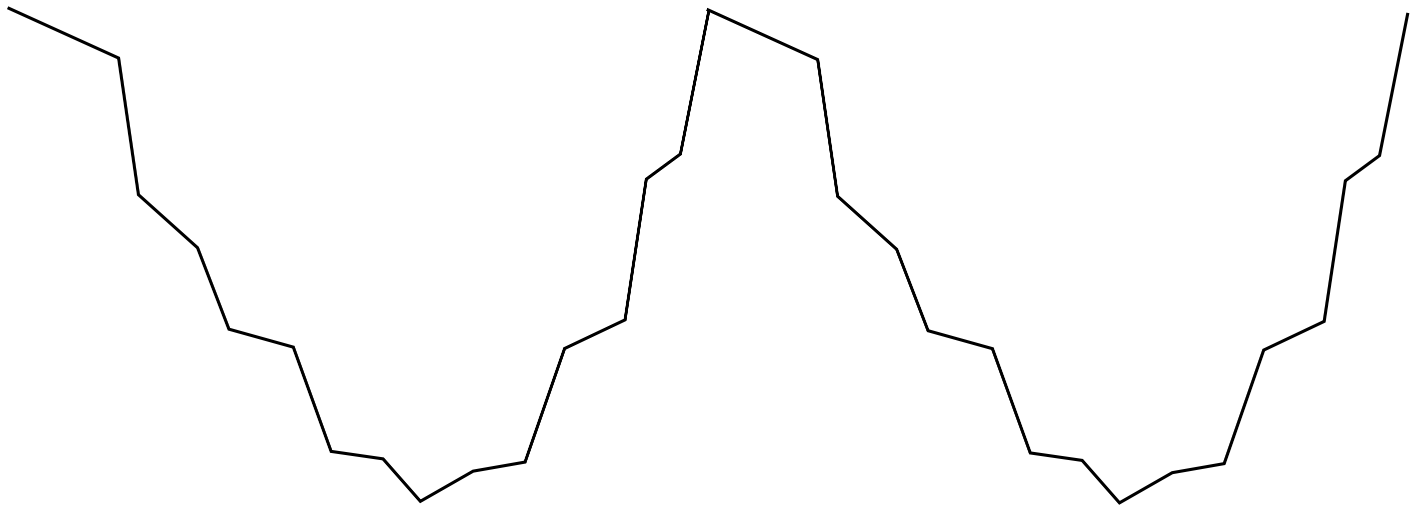
- We first address how overparametrization affects the energy landscapes $E(\Theta), \hat{E}(\Theta)$.
- **Goal 1**: Study simple *topological* properties of these landscapes for half-rectified neural networks.
- **Goal 2**: Estimate simple *geometric* properties with efficient, scalable algorithms. Diagnostic tool.

Non-convexity ≠ Not optimizable



- We can perturb any convex function in such a way it is no longer convex, but such that gradient descent still converges.
- E.g. quasi-convex functions.

Non-convexity $\neq$ Not optimizable



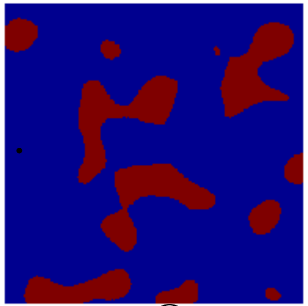
$$F(\theta) = F(g.\theta)\ ,\ g \in G \text{ compact.}$$

- We can perturb any convex function in such a way it is no longer convex, but such that gradient descent still converges.
- E.g. quasi-convex functions.
- In particular, deep models have internal symmetries.

Sublevel sets and topology

- Given loss $E(\theta)$, $\theta \in \mathbb{R}^d$, we consider its representation in terms of level sets:

$$E(\theta) = \int_0^\infty \mathbf{1}(\theta \in \Omega_u) du \ , \ \Omega_u = \{y \in \mathbb{R}^d \ ; \ E(y) \leq u\} \ .$$



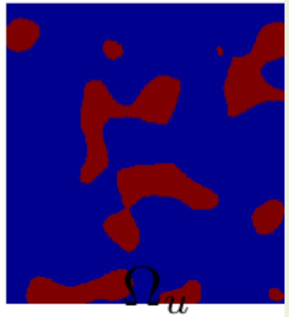
Ω_u

- A first notion we address is about the topology of the level sets .

- In particular, we ask how connected they are, i.e. how many connected components N_u at each energy level u?

Topology of Non-convex Risk Landscape

- A first notion we address is about the topology of the level sets .
  - In particular, we ask how connected they are, i.e. how many connected components N_u at each energy level u?

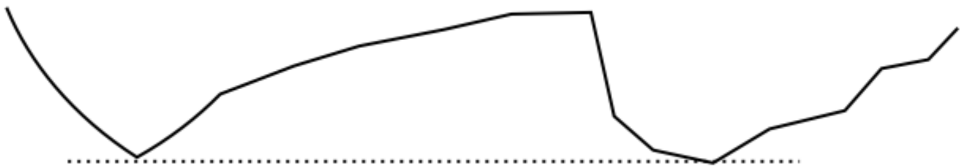


- This is directly related to the question of global minima:

Proposition: If $N_u = 1$ for all u then E has no poor local minima.

$$\text{(i.e. no local minima } y^* \text{ s.t. } E(y^*) > \min_y E(y))$$

- We say E is *simple* in that case.
- The converse is clearly not true.



Weaker: P.1, no spurious local valleys

Given a parameter space Θ and a loss function $L(\theta)$ as in (2), for all $c \in \mathbb{R}$ we define the sub-level set of L as

$$\Omega_L(c) = \{\theta \in \Theta : L(\theta) \leq c\}.$$

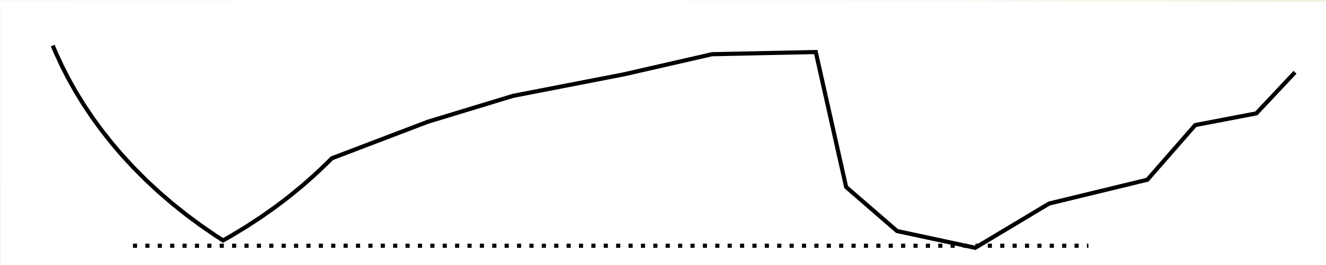
We consider two (related) properties of the optimization landscape. The first one is the following:

P.1 Given any *initial* parameter $\theta_0 \in \Theta$, there exists a continuous path $\theta : t \in [0,1] \mapsto \theta(t) \in \Theta$ such that:

(a) $\theta(0) = \theta_0$

(b) $\theta(1) \in \arg\min_{\theta \in \Theta} L(\theta)$

(c) The function $t \in [0,1] \mapsto L(\theta(t))$ is non-increasing.



Deep Linear Networks

- Some authors have considered linear "deep" models as a first step towards understanding nonlinear deep models:

$$E(W_1, \dots, W_K) = \mathbb{E}_{(X,Y)\sim P} \|W_K \dots W_1 X - Y\|^2 \ .$$

$$X \in \mathbb{R}^n \ , \ Y \in \mathbb{R}^m \ , \ W_k \in \mathbb{R}^{n_k \times n_{k-1}} \ .$$

Theorem: [Kawaguchi'16] If $\Sigma = \mathbb{E}(XX^T)$ and $\mathbb{E}(XY^T)$ are full-rank and Σ has distinct eigenvalues, then $E(\Theta)$ has no poor local minima.

- studying critical points.
- later generalized in [Hardt & Ma'16, Lu & Kawaguchi'17]

Overparametric DLN -> Simple connectivity

$$E(W_1, \ldots, W_K) = \mathbb{E}_{(X,Y)\sim P} \|W_K \ldots W_1 X - Y\|^2 .$$

Proposition: [BF'16]

1. If $n_k > \min(n, m)$, $0 < k < K$, then $N_u = 1$ for all u.
2. (2-layer case, ridge regression) $E(W_1, W_2) = \mathbb{E}_{(X,Y)\sim P} \|W_2 W_1 X - Y\|^2 + \lambda(\|W_1\|^2 + \|W_2\|^2)$ satisfies $N_u = 1 \; \forall \; u$ if $n_1 > \min(n, m)$.

- We pay extra redundancy price to get simple topology.

$$E(W_1, \ldots, W_K) = \mathbb{E}_{(X,Y)\sim P} \|W_K \ldots W_1 X - Y\|^2 \; .$$

Proposition: **[BF'16]**

1. If $n_k > \min(n, m)$, $0 < k < K$, then $N_u = 1$ for all u.
2. (2-layer case, ridge regression)
   $E(W_1, W_2) = \mathbb{E}_{(X,Y)\sim P} \|W_2 W_1 X - Y\|^2 + \lambda(\|W_1\|^2 + \|W_2\|^2)$
   satisfies $N_u = 1 \; \forall \; u$ if $n_1 > \min(n, m)$.

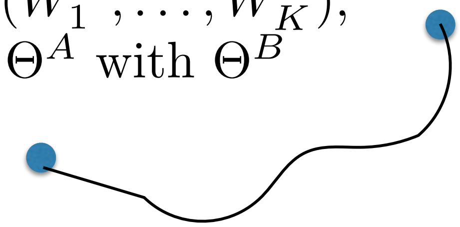
- We pay extra redundancy price to get simple topology.
- This simple topology is an "artifact" of the linearity of the network:

Proposition: **[BF'16]** For any architecture (choice of internal dimensions), there exists a distribution $P_{(X,Y)}$ such that $N_u > 1$ in the ReLU $\rho(z) = \max(0, z)$ case.

Proof Sketch

- Goal:

Given $\Theta^A = (W_1^A, \dots, W_K^A)$ and $\Theta^B = (W_1^B, \dots, W_K^B)$, we construct a path $\gamma(t)$ that connects Θ^A with Θ^B st $E(\gamma(t)) \leq \max(E(\Theta^A), E(\Theta^B))$.



- Main idea:

1. Induction on K.
2. *Lift* the parameter space to $\widetilde{W} = W_1 W_2$: the problem is convex $\Rightarrow$ there exists a (linear) path $\widetilde{\gamma}(t)$ that connects Θ^A and Θ^B.
3. Write the path in terms of original coordinates by *factorizing* $\widetilde{\gamma}(t)$.

- Simple fact:

If $M_0, M_1 \in \mathbb{R}^{n \times n'}$ with $n' > n$,
then there exists a path $t : [0,1] \to \gamma(t)$
with $\gamma(0) = M_0$, $\gamma(1) = M_1$ and
$M_0, M_1 \in \operatorname{span}(\gamma(t))$ for all $t \in (0,1)$.

Group Symmetries

[with L. Venturi, A. Bandeira, '17]

- Q: How much extra redundancy are we paying to achieve $N_u = 1$ instead of simply no poor-local minima?

  – In the multilinear case, we don't need $n_k > \min(n, m)$

  ❖ We do the same analysis in the quotient space defined by the equivalence relationship $W \sim \tilde{W} \Leftrightarrow W = \tilde{W}U \ , \ U \in GL(\mathbb{R}^n) \ .$

Corollary [LBB'17]: The Multilinear regression $\mathbb{E}_{(X,Y)\sim P}\|W_1 \dots W_k X - Y\|^2$ has no poor local minima.

❖ Construct paths on the Grassmanian manifold of subspaces.

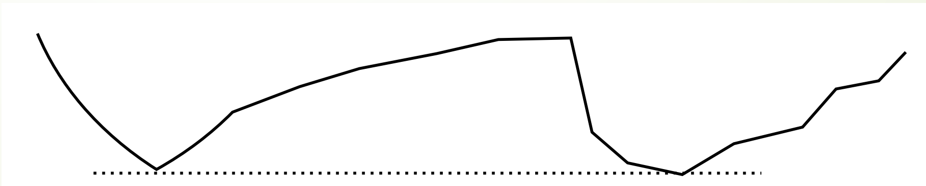
❖ Generalizes best known results for multilinear case (no assumptions on data covariance).

Venturi-Bandeira-Bruna'18

$$\Phi(x;\theta) = W_{K+1} \cdots W_1 x \ , \tag{13}$$

where $\theta = (W_{K+1}, W_K, \ldots, W_2, W_1) \in \mathbb{R}^{n \times p_{K+1}} \times \mathbb{R}^{p_{K+1} \times p_K} \times \cdots \mathbb{R}^{p_2 \times p_1} \times \mathbb{R}^{p_1 \times n}$.

Theorem 8 *For linear networks* (13) *of any depth $K \geq 1$ and of any layer widths $p_k \geq 1$, $k \in [1, K+1]$, and input-output dimensions n, m, the square loss function* (2) *admits no spurious valleys.*



Asymptotic Connectedness of ReLU

- Good behavior is recovered with nonlinear ReLU networks, provided they are sufficiently overparametrized:
- Setup: two-layer ReLU network:

$$\Phi(X;\Theta) = W_2\rho(W_1 X)\ ,\ \rho(z) = \max(0,z). W_1 \in \mathbb{R}^{m\times n}, W_2 \in \mathbb{R}^m$$

Theorem [BF'16]: For any $\Theta^A, \Theta^B \in \mathbb{R}^{m\times n}, \mathbb{R}^m$, with $E(\Theta^{\{A,B\}}) \leq \lambda$, there exists path $\gamma(t)$ from Θ^A and Θ^B such that $\forall\ t\ , E(\gamma(t)) \leq \max(\lambda, \epsilon)$ and $\epsilon \sim m^{-\frac{1}{n}}$.

- Overparametrisation "wipes-out" local minima (and group symmetries).
- The bound is cursed by dimensionality, ie exponential in n .
- Result is based on local linearization of the ReLU kernel (hence exponential price).

Asymptotic Connectedness of ReLU

- Good behavior is recovered with nonlinear ReLU networks, provided they are sufficiently overparametrized:
- Setup: two-layer ReLU network:

$$\Phi(X;\Theta) = W_2\rho(W_1 X) \ , \ \rho(z) = \max(0, z). W_1 \in \mathbb{R}^{m\times n}, W_2 \in \mathbb{R}^m$$

Theorem [BF'16]: For any $\Theta^A, \Theta^B \in \mathbb{R}^{m\times n}, \mathbb{R}^m$, with $E(\Theta^{\{A,B\}}) \leq \lambda$, there exists path $\gamma(t)$ from Θ^A and Θ^B such that $\forall\ t\ , E(\gamma(t)) \leq \max(\lambda, \epsilon)$ and $\epsilon \sim m^{-\frac{1}{n}}$.

- Overparametrisation "wipes-out" local minima (and group symmetries).
- The bound is cursed by dimensionality, ie exponential in n .
- *Open question:* polynomial rate using Taylor decomp of $\rho(z)$?

Kernels are back?

- The underlying technique we described consists in "convexifying" the problem, by mapping *neural* parameters $\boldsymbol{\Theta}$

$$\Phi(x;\Theta) = W_k \rho(W_{k-1} \dots \rho(W_1 X))) \ , \ \Theta = (W_1, \dots W_k) \ ,$$

to *canonical* parameters $\beta = \mathcal{A}(\Theta)$:

$$\Phi(X;\Theta) = \langle \Psi(X), \mathcal{A}(\Theta) \rangle \ .$$

Kernels are back?

- The underlying technique we described consists in "convexifying" the problem, by mapping *neural* parameters Θ

$$\Phi(x;\Theta) = W_k \rho(W_{k-1} \dots \rho(W_1 X))) \ , \ \Theta = (W_1, \dots W_k) \ ,$$

to *canonical* parameters $\beta = \mathcal{A}(\Theta)$:

$$\Phi(X;\Theta) = \langle \Psi(X), \mathcal{A}(\Theta) \rangle \ .$$

- Second layer setup: $\rho(\langle w, X \rangle) = \langle \mathcal{A}(w), \Psi(X) \rangle$.

Corollary: [BBV'17] If $\dim\{\mathcal{A}(w), w \in \mathbb{R}^n\} = q < \infty$ and $M \geq 2q$, then $E(W,U) = \mathbb{E}|U\rho(WX) - Y|^2$, $W \in \mathbb{R}^{M \times N}$ has no poor local minima if $M \geq 2q$.

VBB'18

Theorem 5 *The loss function*

$$L(\theta) = \mathbb{E}\|\Phi(X;\theta) - Y\|^2$$

of any network $\Phi(x;\theta) = U\rho W x$ *with effective intrinsic dimension* $q < \infty$ *admits no spurious valleys, in the over-parametrized regime* $p \geq q$*. Moreover, in the over-parametrized regime* $p \geq 2q$ *there is only one global valley.*

We notice that the same optimal representation functions $\Phi(\cdot;\theta)$ could also be obtained using a generalized linear model, where the representation function has the linear form $\Phi(x;\theta) = \langle\theta, \varphi(x)\rangle$, with the same underlying family of representation functions $V_{\mathcal{X}}$. A main difference between the two models is that the former requires the choice of a non-linearity, that is of any activation function ρ, while the latter implies the choice of a kernel functions. The non-trivial fact captured by our result

Kernels are back?

- The underlying technique we described consists in "convexifying" the problem, by mapping *neural* parameters Θ

$$\Phi(x;\Theta) = W_k \rho(W_{k-1} \dots \rho(W_1 X))) \ , \ \Theta = (W_1, \dots W_k) \ ,$$

to *canonical* parameters $\beta = \mathcal{A}(\Theta)$

$$\Phi(X;\Theta) = \langle \Psi(X), \mathcal{A}(\Theta) \rangle \ .$$

- This is precisely the formulation of ERM in terms of Reproducing Kernel Hilbert Spaces [Scholkopf, Smola, Gretton, Rosasco, …]
- Recent works developed RKHS for Deep Convolutional Networks
  - [Mairal et al.'17, Zhang, Wainwright & Liang '17]
  - See also F. Bach's talk tomorrow [Bach'15].
  - Open question: behavior of SGD in Θ in terms of canonical params? Progress on matrix factorization, e.g [Srebo'17]

Polynomial Activations

$$\rho(z) = a_0 + a_1 z + \cdots + a_d z^d. \tag{10}$$

In this case, we have:

Corollary 6 *For two-layers NNs $\Phi(x;\theta) = U\rho W x$, if the activation function ρ is of the form (10), then the square loss function* (2) *admits no spurious valleys in the over-parametrized regime*

$$p \geq \sum_{i=1}^{d} \binom{n+i-1}{i} \mathbf{1}_{\{a_i \neq 0\}} = O(n^d). \tag{11}$$

Between linear and ReLU: polynomial nets

- Quadratic nonlinearities $\rho(z) = z^2$ are a simple extension of the linear case, by lifting or "kernelizing":

$$\rho(Wx) = \mathcal{A}_W X \ , \ \ X = xx^T \ , \ \ \mathcal{A}_W = (W_k W_k^T)_{k \leq M} \ .$$

- We have the following extension:

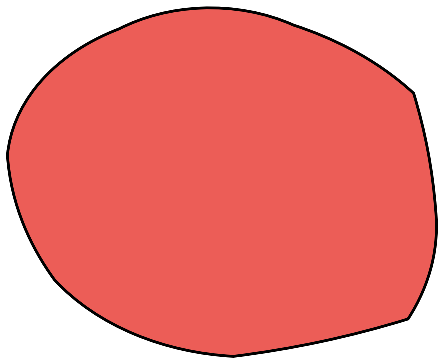
Proposition: If $M \geq 3N^2$, then the landscape of two-layer quadratic network is simple: $N_u = 1 \ \forall \ u$.

Proposition: If $M_k \geq 3N^{2^k} \ \forall \ k \leq K$, then the landscape of K-layer quadratic network is simple: $N_u = 1 \ \forall \ u$.

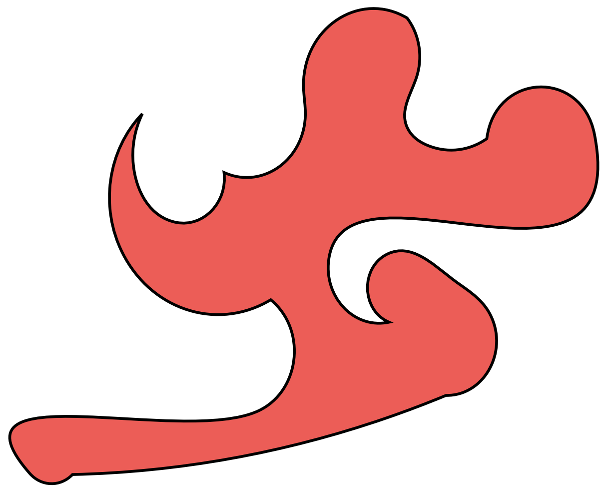
- *Open question:* Improve rate by exploiting Group symmetries? Currently we only win on the constants.

From Topology to Geometry

- The next question we are interested in is conditioning for descent.
- Even if level sets are connected, how easy it is to navigate through them?
- How "large" and regular are they?



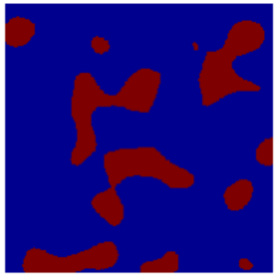
easy to move from one energy level to lower one



hard to move from one energy level to lower one

Finding Connected Components

- Suppose $\theta_1,\ \theta_2$ are such that $E(\theta_1) = E(\theta_2) = u_0$
  - They are in the same connected component of Ω_{u_0} iff there is a path $\gamma(t),\ \ \gamma(0) = \theta_1, \gamma(1) = \theta_2$ such that
    $$\forall\ t \in (0,1)\ ,\ E(\gamma(t)) \leq u_0\ .$$



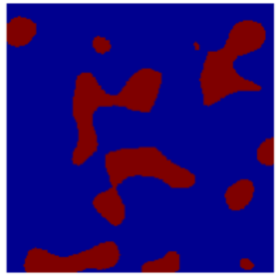
Ω_u

  - Moreover, we penalize the length of the path:

$$\forall\ t \in (0,1)\ ,\ E(\gamma(t)) \leq u_0 \quad \text{and} \quad \int \|\dot{\gamma}(t)\| dt \leq M\ .$$

Finding Connected Components

- Suppose θ_1, θ_2 are such that $E(\theta_1) = E(\theta_2) = u_0$
  - They are in the same connected component of Ω_{u_0} iff there is a path $\gamma(t)$, $\gamma(0) = \theta_1, \gamma(1) = \theta_2$ such that $\forall\ t \in (0,1)\ ,\ E(\gamma(t)) \leq u_0\ .$



Ω_u

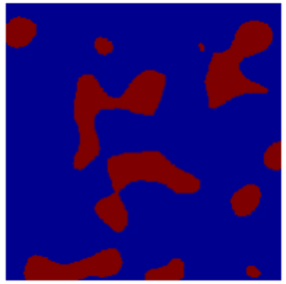
  - Moreover, we penalize the length of the path:

$$\forall\ t \in (0,1)\ ,\ E(\gamma(t)) \leq u_0 \ \text{ and } \int \|\dot{\gamma}(t)\| dt \leq M\ .$$

- Dynamic programming approach:

Finding Connected Components

- Suppose $\theta_1,\ \theta_2$ are such that $E(\theta_1) = E(\theta_2) = u_0$
  - They are in the same connected component of Ω_{u_0} iff there is a path $\gamma(t),\ \gamma(0) = \theta_1, \gamma(1) = \theta_2$ such that $\forall\ t \in (0,1)\ ,\ E(\gamma(t)) \leq u_0\ .$



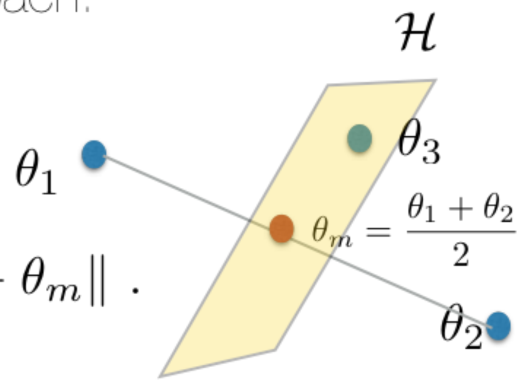
Ω_u

  - Moreover, we penalize the length of the path:

$$\forall\ t \in (0,1)\ ,\ E(\gamma(t)) \leq u_0 \ \text{ and } \int \|\dot{\gamma}(t)\| dt \leq M\ .$$

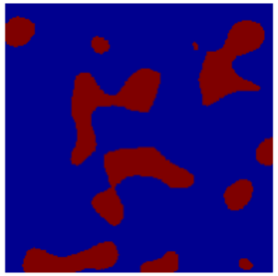
- Dynamic programming approach:

$$\theta_3 = \arg \min_{\theta \in \mathcal{H};\, E(\theta) \leq u_0} \|\theta - \theta_m\|\ .$$



Finding Connected Components

- Suppose $\theta_1,\ \theta_2$ are such that $E(\theta_1) = E(\theta_2) = u_0$
  - They are in the same connected component of Ω_{u_0} iff there is a path $\gamma(t),\ \ \gamma(0) = \theta_1, \gamma(1) = \theta_2$ such that $\forall\ t \in (0,1)\ ,\ E(\gamma(t)) \leq u_0\ .$



Ω_u

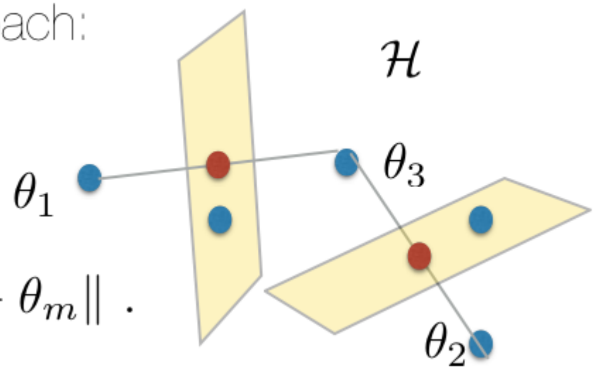
  - Moreover, we penalize the length of the path:

$$\forall\ t \in (0,1)\ ,\ E(\gamma(t)) \leq u_0\ \ \text{and}\ \int \|\dot{\gamma}(t)\| dt \leq M\ .$$

- Dynamic programming approach:

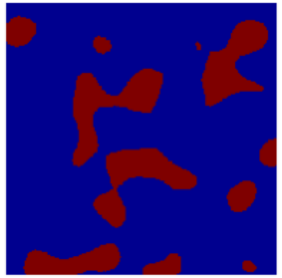
$$\theta_m = \frac{\theta_1 + \theta_2}{2}$$

$$\theta_3 = \arg \min_{\theta \in \mathcal{H};\, E(\theta) \leq u_0} \|\theta - \theta_m\|\ .$$



Finding Connected Components

- Suppose $\theta_1,\ \theta_2$ are such that $E(\theta_1) = E(\theta_2) = u_0$
  - They are in the same connected component of Ω_{u_0} iff there is a path $\gamma(t),\ \gamma(0) = \theta_1, \gamma(1) = \theta_2$ such that $\forall\ t \in (0,1)\ ,\ E(\gamma(t)) \leq u_0\ .$



Ω_u

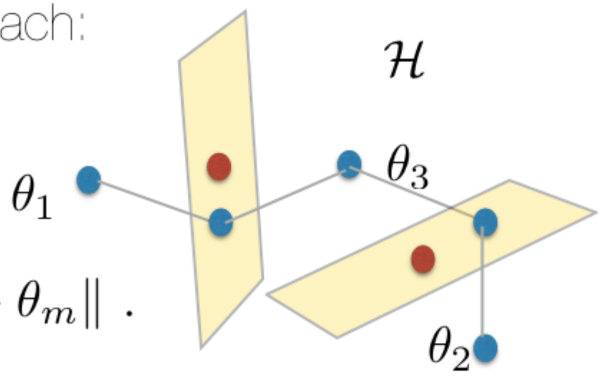
  - Moreover, we penalize the length of the path:

$$\forall\ t \in (0,1)\ ,\ E(\gamma(t)) \leq u_0 \ \text{ and } \int \|\dot{\gamma}(t)\| dt \leq M\ .$$

- Dynamic programming approach:

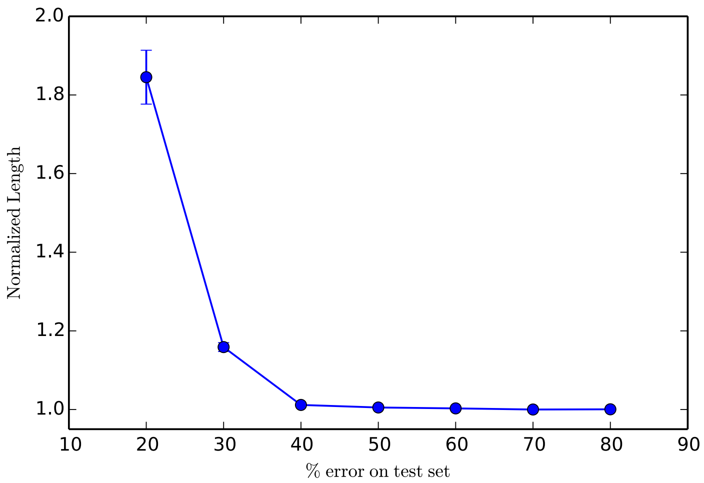
$$\theta_m = \frac{\theta_1 + \theta_2}{2}$$

$$\theta_3 = \arg \min_{\theta \in \mathcal{H};\, E(\theta) \leq u_0} \|\theta - \theta_m\|\ .$$

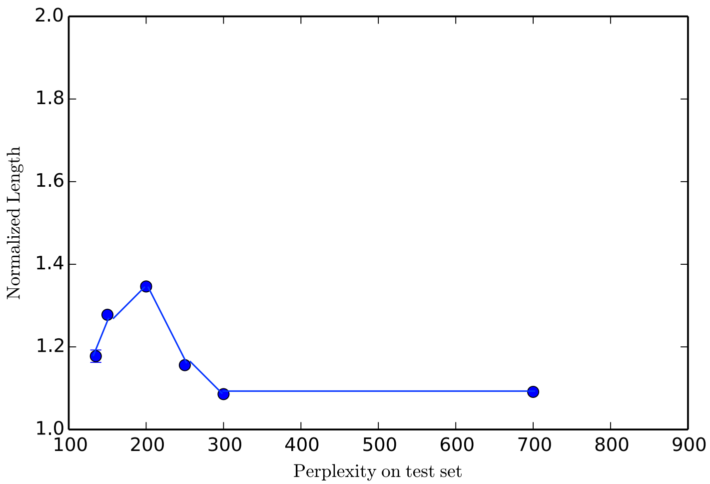


Numerical Experiments

- Compute length of geodesic in Ω_u obtained by the algorithm and normalize it by the Euclidean distance. Measure of curviness of level sets.



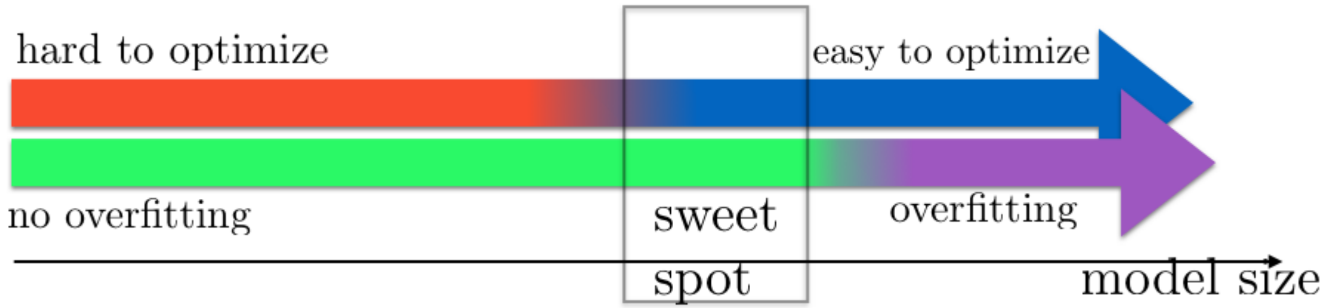
CNN/CIFAR-10



LSTM/Penn

Analysis and perspectives

- #of components does not increase: no detected poor local minima so far when using typical datasets and typical architectures (at energy levels explored by SGD).
- Level sets become more irregular as energy decreases.
- Presence of "energy barrier"?
- Kernels are back? CNN RKHS
- Open: "sweet spot" between overparametrisation and overfitting?
- Open: Role of Stochastic Optimization in this story?



Summary

- **Overparameterization may lead to simple risk landscapes with flat global minima**
- **GD/SGD may find flat global minima**
- GD may find max margin global minima

Thank you!